The Application of Semantic Technologies to Scientific Archives

J. Steven Hughes
Daniel J. Crichton

Science Archives in the 21st Century
NSSDC, Washington D.C.
25-26 April 2007

steve.hughes@jpl.nasa.gov

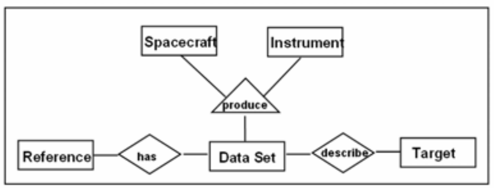
Background

- The Information Model is the foundation on which an information system is built.
  - The information model identifies and defines:
    - the elements to be processed
    - the entities that provide context for the elements
    - the relationships that provide meaning for the elements
  - It evolves at a speed different from and outlasts any information technology choice
  - It is best developed and maintained independent of all implementation choices
    - This includes the languages into which it itself will be implemented
- The development and the subsequent management of the Information Model is the most significant factor for developing successful information systems on time, within budget, and that remain viable.
  - This is especially true for the development of science information systems where the science environment continually evolves.
  - For example, Science Archives must manage ever more complex data collected from increasingly sophisticated instrumentation

The Issue

- The information model is seldom developed or managed over time with the rigor afforded to software development and management. Consider the Planetary Data System as an example where this has been attempted.

1990
Captured using E-R Model



→

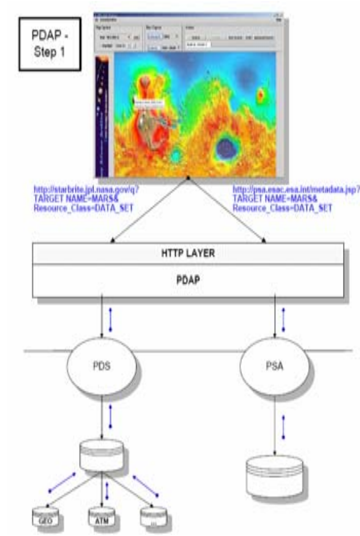
2007
17 years of evolutionary changes; 3 languages; 3 technologies

```
OBJECT            = INSTRUMENT
  INSTRUMENT_ID   = VISA
  SCID            = VO1
  INSTRUMENT_NAME = VISUAL_IMAGING...
  INSTRUMENT_TYPE = VIDICON_CAMERA
  ...
END_OBJECT
```

→

2007+
International Planetary Data Alliance (IPDA) is adopting the PDS Information Model

10 member agencies



1. ***How is a data model to be maintained over time, independent of its implementation***
2. ***How is interoperability between data models to be managed***

Problem Definition

- Require an information model development and management framework that includes methodologies and tools to address the following tasks across geographical, political, and domain boundaries.
    1. Develop archive information models for diverse and complex science domains
    2. Manage continuous evolutionary changes and additions to the information model.
    3. Maintain the information model independent of system implementation choices.
    4. Use the information model to drive implementations and documentation.
    5. Enable and maintain interoperability between domains.

Requirements

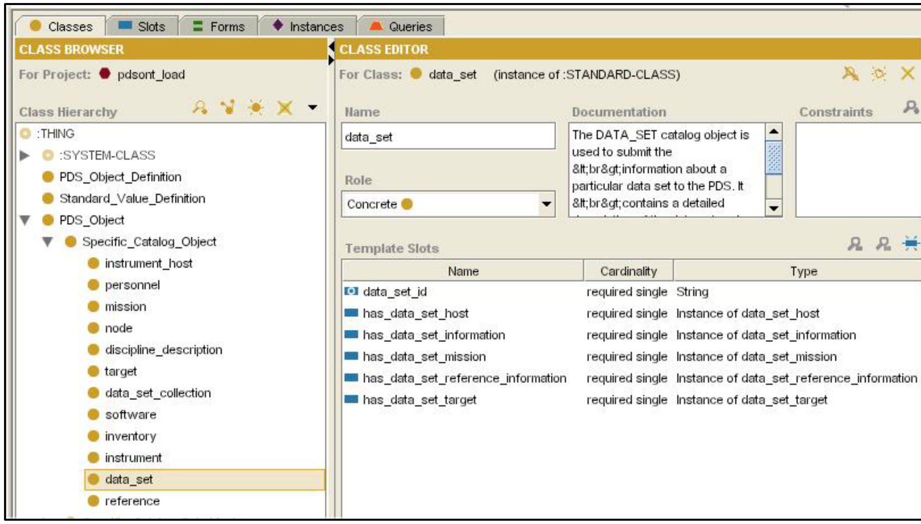
- Independent - Exist as an independent, comprehensive source for the information model
- Scalable – Allow increase in size, sophistication, and complexity of the model.
- Evolvable – Allow model to change as the domain changes.
- Flexible – Allow translation to diverse notations and languages for documentation, communication, and implementation purposes.
- Capable - Maintain the upper bound on semantic richness for all domain information needs
- Accessible - Provide open access for users, developers, and systems.
- Usable – Make model information available for use and processing by both machines and humans

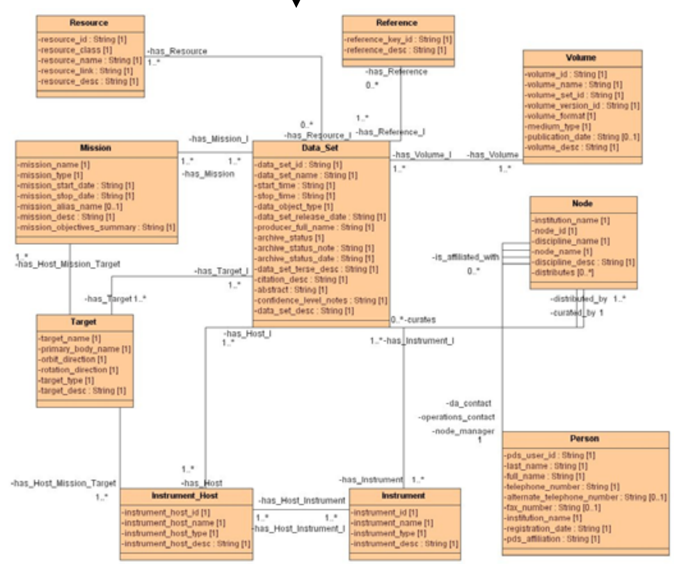
Proposal

- Ontology Modeling Tool
  - Captures the information resulting from answering the “what is ___?” question.
    - The ontological answer to “what is a Viking image” is a collection of metadata that formally describes a Viking image sufficiently for some stated purpose.
  - Typically includes the definition of classes and attributes and their relationships.
  - Metadata is stored in a framework that is expressible as a notation or language. (e.g. Frames, RDF, OWL)
  - Metadata can be exported to other notations or languages. (e.g. RDF, N3, OWL, XMI)
- Ingesting data into the ontology results in a knowledge base.
  - Tests the ontology.
  - Simple, full-capability “database” applications can easily be configured.
- The modeling information can easily be exported to traditional data modeling notations and tools for implementation and documentation purposes (e.g. UML)
- Semantic technologies can process and reason about the model
  - “Triple” database engines provide both unfettered and schema guided access to both data and metadata.
  - Semantic browsers provide both text- and facet-base search and guided navigation through the knowledge base.
  - Development frameworks are available for configuring custom GUI applications.
  - Semantic technologies such as “reasoners” provide sophisticated analysis

PDS Case Study

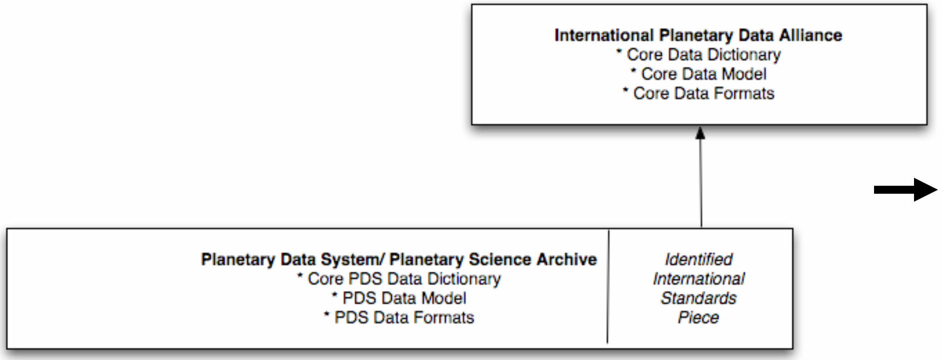
- Upper level model (PDS Catalog)
  - Captures Data Sets, Instruments, Missions, etc
  - Model information extracted from data dictionary
  - Hardcopy documentation used to complete the model and add descriptive narrative
- Product Model
  - Captures basic data objects as extracted from data dictionary. (Image, TimeSeries, etc)
  - Captures product types as extracted from products in the archive.
  - PDS technical staff identify “core” product types.





International Planetary Data Alliance (IPDA)

- IPDA Archive Data Standards Requirements Identification Project
  - Identify the existing subset of standards used by PDS/PSA which are appropriate for internationalization
  - Compile a draft Information Model
    - Data Dictionary of Terms
    - Standard Data Formats
    - Model of Objects and their Relationships
  - Generate Information Model Document from PDS ontology
    - Export “Core” elements to RDF/XML
    - Generate LaTEX file.



2.2.4 INSTRUMENT

Object Type: Specific_Catalog_Object
Object Description: The INSTRUMENT catalog object is used to submit information to the PDS about a particular instrument. It contains a detailed description and identifies any references associated with the instrument.

Relationship	Entity	Card
has_Data_Element	instrument_desc	1
	instrument_host_id	1
	instrument_id	1
	instrument_name	1
	instrument_type	1
	reference_key_id	1..*